Review of: "NER Sequence Embedding of Unified Medical Corpora to Incorporate Semantic Intelligence in Big Data Healthcare Diagnostics"

Claudia Lanza[1]

1 University of Calabria

Potential competing interests: No potential competing interests to declare.

The paper presents a study targeted to the analysis of diabetes mellitus (DM) and its comorbidity diseases through NER sequence embedding approaches using advanced ML integrated with NLP techniques. The objectives of this work aren't well highlighted in the abstract as well as in the introduction, therefore the authors should stress a little bit more which are the final perspectives of their work as mentioned in the title of their paper.

I also suggest to split the Related work section in subparagraphs according to the topic addressed, sometimes it seems difficult to read it because there are so many subjects included.

In the Related work section I feel to suggest authors to consider the inclusion of the following published scientific works on the semantic annotation and Neural word emeddings within the medicine field of study:

- Pasceri, Erika, Mérième Bouhandi, Claudia Lanza, Anna Perri, Valentina Laganà, Raffaele Maletta, Raffaele Di Lorenzo, and Amalia C. Bruni. 2023. “Neurodegenerative Clinical Records Analyzer: Detection of Recurrent Patterns Within Clinical Records towards the Identification of Typical Signs of Neurodegenerative Disease History”. *JLIS.It* 14 (2):20-38. https://doi.org/10.36253/jlis.it-522.;
- Attardi, Giuseppe, Vittoria Cozza and Daniele Sartiano. “Annotation and Extraction of Relations from Italian Medical Records.” *Italian Information Retrieval Workshop* (2015).;
- Wu, Yonghui, Jun Xu, Min Jiang, Yaoyun Zhang, e Hua Xu. 2015. «A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text». AMIA ... Annual Symposium Proceedings. AMIA Symposium 2015: 1326–33.;
- Co-Morbidity and Smoking Status for Asthma Research: Evaluation of a Natural Language Processing System». BMC Medical Informatics and Decision Making 6 (luglio): 30. https://doi.org/10.1186/1472-6947-6-30. DOI: https://doi.org/10.1186/1472-6947-6-30

Images are sometimes inverted, it's hard to read the info in them.

In the sub-section Word Embedding for Clinical Notes and Practitioner Comments, what do the authors mean by saying “the vocabulary size”? An explanation could be useful.

I also suggest to specify in Figure 6 what is the “importance” of the words for the authors.

In section 4.3. a bibliographic reference is missing when authors write “The quality of embedding models is based on the size and type of corpus whether general or domain-specific. In a large general corpus, there is a large vocabulary to infer. The domain-specific corpus is inferred for semantic similarity of terms as in our case of clinical diagnoses.”

A proofreading is recommended mostly for some typos.